3F－5

UNIXワークステーションによる高可用性システムのデータ複製機構

岡田　英明，坂倉　隆史，上村　ホゼ，菅　隆志

三菱電機(株) 情報システム研究所

1　はじめに

近年のコンピュータのダウンサイジングの流れの中で、UNIXワークステーション上においてデータベース等のビジネスアプリケーションを利用する傾向が増大しつつある。このようなビジネス分野での利用においてはシステムの信頼性、可用性が強く求められるが、一般にはUNIXワークステーションシステムはそれに応えられるものではない。

システムの信頼性を高める手段として、ハードウェアに冗長性を持たせたフォルトトレラントシステムを利用する方法がある。このようなシステムは専用のハードウェアを必要とすることなどにより高価なものとなる。一方で、冗長性の付加を複数のワークステーションを高速なネットワークで接続することによって実現する方法について研究、開発が行なわれている。個々のワークステーションの信頼性が低くとも、システム全体ではデータベース等のサービスを高可用度で提供する。このような高可用性システムは、より安価な高信頼システムの一形態として期待されている。

我々はこのような、特殊なハードウェアを使用しない高可用性システムの構築を考案中である。本論文では、高可用性システムの一部であるネットワークを用いたデータ複製機構について提案する。

2　高可用性システムの概要

我々の考える高可用性システムとは、図1に示すようなシステムである。前節でも述べたように、各ノードは標準的なワークステーションまたはPCであり、OS(UNIX)、アプリケーションも標準的なものを使用することを前提とする。

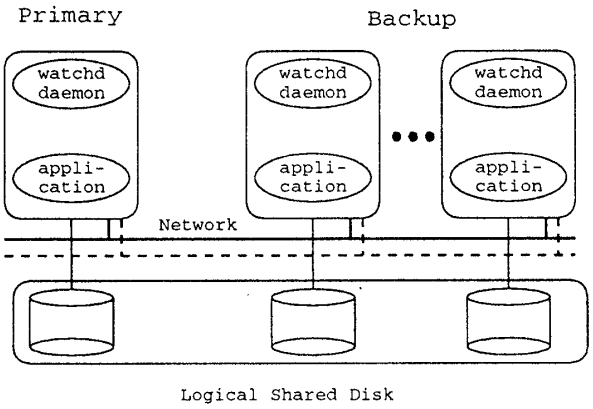

図 1: 高可用性システムの構成

1つの稼働系 (Primary Node) に対し、1つ以上の待機系 (Backup Node) を持つ。各ノードにはシステムの状態を監視する機構が備えられている。この監視機構が稼働系において何らかの障害を検知し、その障害が即時には回復不可能である場合、系の切替えを行なう。待機系の中の1ノードを新たな稼働系とし、中断されたサービスを提供する。

新しく稼働系となったノードが中断されたサービスを提供するためには、サービスを提供するための情報の他に、元の稼働系が行なっていたサービスの状態や履歴などの情報を必要とする。この情報を得るために稼働系と待機系に論理的な共有記憶装置を持つ。稼働系、待機系のそれぞれ物理的には異なる記憶装置が同じデータを持ち、サービス実行中におけるデータの更新に対してはデータ複製機構を利用することによりデータの共有化を実現する。

3　データ複製機構の概要

ファイルシステムにおける上に述べたような共有記憶装置の実現が行なわれているが[1]、我々は、

Data Replication Mechanism for a Commodity Based High Availability UNIX System

Hideaki Okada, Takashi Sakakura, Joe Uemura, Takashi Kan

Computer and Information Systems Laboratory. Mitsubishi Electric Corp.

より汎用的なデバイスレベルにおいて、いくつかのデータ複製機構を考案している。今回、提案する機構はその中の１つであり、その構成を待機系が１つの場合について図2に示す。特殊なハードウェアは使用しておらず、UNIX の稼働する２台のワークステーションおよびイーサネットなどのネットワーク、SCSI ディスク等の外部記憶装置からなる。これらにソフトウェアとして、Logical Volume Manager、Network Disk を付加する。これらはデバイスドライバレベルで実現されるものであり、ドライバ以外のカーネルへの修正は必要としない。

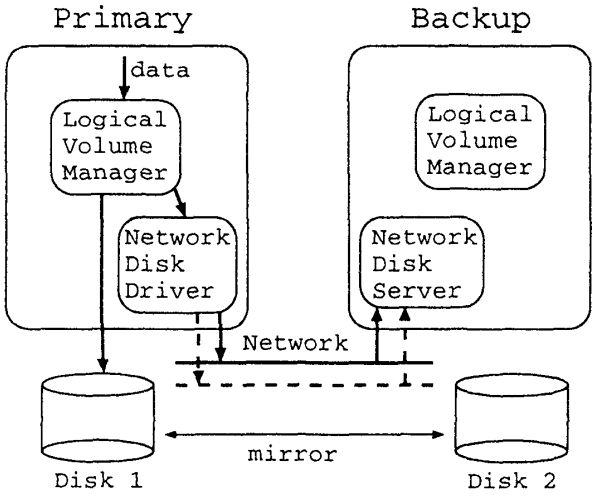

図 2: データ複製機構の構成

Network Disk Driver は、疑似デバイスドライバであり、稼働系における Network Disk のデバイスに対する読み書きをネットワークを介して待機系の Network Disk Server へ送り、Network Disk Server は待機系ノードのデバイス (Disk2) へのアクセスを行なう。

Logical Volume Manager は、疑似デバイスドライバを含む、ディスクを管理するソフトウェアであり、このデータ複製機構では Logical Volume Manager のディスクミラーリング機能を使用する。ミラーリング機能を有する Logical Volume Manager の疑似デバイスドライバに対する書き込みに対し、このドライバはその書き込みを複数のデバイスに同じように行ない、データの複製を行なう。データの読み込みに際しては、１つのデバイスからのみ行なう、双方のデバイスから交互に行なうなど、状況に応じて使い分ける。

以上、２つのソフトウェアを組み合わせることにより、稼働系におけるデータの書き込みに対し、Disk1, Disk2 の双方へのデータの書き込みが可能となる。データを読み込む時には、性能を考慮して Disk1 のみから読み込む。そして、稼働系に障害が発生した時には、Disk2 を待機系の Logical Volume Manager の管理下におき、これを利用してサービスを提供する。

このようなデータ複製方法は、既に述べたようにデバイスドライバレベルのソフトウェアの付加のみで実現できる、汎用的なものである。このような利点を持つ一方、通常のディスクアクセス時間に加えてネットワークを使用する時間がかかるため性能面での問題がある。しかし、ATM 等の高速ネットワーク網が実現すれば性能はかなり改善されることが予想される。また、ネットワークを用いていることから、距離の大きくはなれた２地点間においても利用が可能であり、ATM 等を利用した WAN へも適用可能である。

4 おわりに

高可用性システムのためのデータ複製機構を提案した。本機構は、特殊なハードウェアを必要とせず、デバイスドライバレベルで実現可能であり、カーネルの修正も要しない、という特徴を持つ。

現在、本データ複製機構、稼働系、待機系の各ノードを監視し、システムスイッチを行なう機構を含め、高可用性システムを試作中である。

今後は、データ複製機構のイーサネットを用いた場合の性能測定、ATM 等の高速ネットワークを用いた場合の性能予測等を行なう予定である。

参考文献

[1] G.Flowler, Y.Huang, D.Korn, H.Rao A User-Level Replicated File System, In *Proceedings of Summer USENIX*, July 1993.